文 章 编 号: 1000-1220(2000) 04-0375-04

基于知识的网页检索工具

廖明宏　吴翔虎
(哈尔滨工业大学计算机科学与工程系　哈尔滨 150001)

摘　要: 随着因特网在全球范围的广泛使用,越来越多的人们借助于因特网从事科研和商务活动,而网页检索工具成了人们必不可少的软件工具.然而,目前流行的检索工具大多基于关键字查询,常常出现信息过载或有用信息丢失等现象.造成这一原因主要有两方面:用户提交的查询不能很好地表达他的目的;查询的结果没有建立有效的索引机制,引导人们快速找到有用信息.为此我们提出一种基于知识的网页检索工具(KWSE),它是在已有的检索工具的基础上增加查询概念和文档改写两项功能,即用户提交的不是关键字而是为达到某一目的的概念;并对返回的文档建立新的索引,指向有用信息.实验表明,这种检索工具更符合人们的习惯,其查询结果更为有效.

关 键 词: 因特网; WWW; 信息检索
分 类 号: TP393.4　　文献标识码: A

1　引　言

近年来,随着因特网在全球范围内的普及和应用,万维网(WWW)成为最重要的电子信息资源.它包含着你所能想象到的各种知识信息.通过多媒体的应用,HTML 文档以不同的形式提供各种有用的知识(如文本、图形、动画、声音、虚拟现实等等).网页文档之间的超文本链将不同的知识单元建立起联系.基于 HTML 标准,用户可利用网页浏览器查看各种资料,并通过 HTML 链访问分布于全球的各种信息和知识单元.然而,从网页上检索有用信息并没有得到很好的解决,常常听到有人抱怨,借助通用的网页检索工具去检索信息,经常得到大量的无用信息,与此同时,他们很难确定哪些是有用的信息.事实上,网页所提供的查询服务大多是基于关键字的检索机制.为了避免信息丢失,大多检索工具给出尽可能多的匹配结果(包括部分匹配的结果).过多的匹配结果给用户带来时间和金钱的浪费,因此,高效的信息检索工具的研制成了一个重要的研究焦点.

我们认为,造成这一问题的主要原因有两点:首先,用户给出的查询关键字不能很好地反映用户的意图;其次,返回的查询结果不能引导用户很快地定位到他们所感兴趣的信息单元.鉴于此,本文介绍的基于知识的网页检索工具(Knowledge-Based Web Search Engine KWSE)是在已有的检索工具的基础上增加概念查询和文档改写两大功能模块而形成的.它是基于客户—服务器模型.为使用户仍能使用他们所熟悉的网页浏览器等软件工具,在客户端不做大的改动.而在服务器一端,我们在常用的网页检索工具(如 Yahoo, Atta Vista 等)增加两项功能:一方面,允许用户以概念形式输入查询要求,系统在知识库的支持下将概念转化为基本关键字提交给检索工具;另一方面,服务器返回的结果文档做进一步的分析和改写,增加一些有用的索引链,帮助用户快速找到他们所感兴趣的信息.

本文在澄清数据、信息和知识这三个基本概念的基础上,介绍了 KWSE 的体系结构;然后重点讨论概念查询的转换和结果文档的分析与改写,最后对该工具的应用前景做了展望.

2　数据、信息和知识

在介绍具体的网页检索工具之前,有必要对数据、信息和知识三个基本概念做一下澄清说明.我们这里不想从哲学、逻辑学的角度对这三个概念做分析和讨论,而是引用 Nygard[1] 和 Erdmann[2] 等人的观点从符号学的三维空间:即语法、语义和语用加以分析.事实上,从纯粹的符号表示上是无法区别数据、信息和知识的,只有通过关系,这些符号才能区分出数据、信息和知识来.

(1) 符号间的关系,即句子的语法,它不涉及到符号与现实世界之间的关系,因此符号在这一维上可看成是数据.

(2) 符号与符号含义间的关系,即涉及到符号的语义.只有通过符号与它们的含义这一层关系,符号才能从数据变成信息.

(3) 符号与它的用户间的关系.在这一维上符号与使用它的用户建立起联系.当用户为实现某一目的而使用这些符号,这个目的称为符号模式或符号知识.这一层关系就是符号的语用属性,它定义了知识的一个重要特性,即只有知识才能使得用户执行动作或做出决策成为可能.

比如,数字"8250"没有赋予它任何含义,它仅仅是数据而已.但"8250 元"即在数字后加上"元",它说明 8250 是一笔钱数.这时它已从数据上升到信息.对于这一信息不同人有不同的反应,例如有一个人想买一台计算机,查询的价格都在 10000 元以上,只有一个公司出售的计算机价格为 8250 元.且技术参数

收稿日期: 1999-03-16　作者简介: 廖明宏,博士,副教授,研究方向为人工智能、并行与分布式处理.吴翔虎,讲师,研究方向为分布式处理、操作系统.
Email: liao@hawk.hit.edu.cn

与其它计算机一样. 很显然, 他会毫不犹豫地购买这台计算机. 可以注意到 "8250 元"在特定的环境中由信息变成了知识.

不论是数据库、信息系统、基于知识的系统或任何其它的计算机系统, 它们所表示的对象都是一样的, 即都表示成符号. 只有通过这些符号的使用, 包括它们的不同作用, 上下文和用户等, 它们才成为数据、信息和知识.

通常, 信息检索本身不是目的. 相反, 每个人查找信息都带有一定的目的性, 他们希望找到的信息能够帮助他们达到这个目的. 因此我们认为, 网页检索的结果应该是有用的知识、而不是毫无关系的信息.

3 KWSE 的体系结构

KWSE 的体系结构是基于客户—服务器模型的(见图 1) .

网页最大的优势就是可被多种操作系统上的浏览器所访问, 为保持这一优势, 让用户可继续使用他们所喜欢的软件(如 Netscape, Internet Explorer 等) , 我们增加的新功能不直接集成到客户端, 而是放在服务器上.

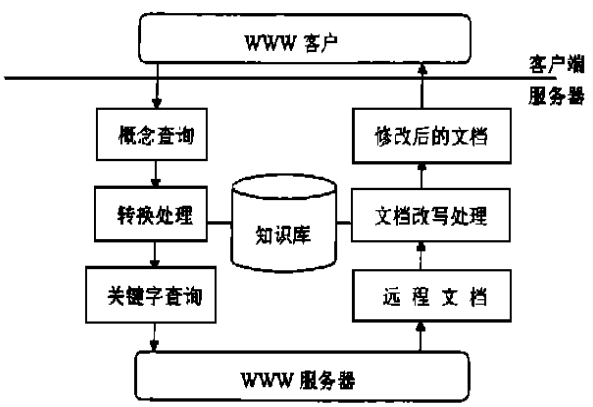

图 1 KWRE 的体系结构

3.1 概念查询与文档改写

检索工具查询可以很复杂, 为使查询的形式更适合用户的习惯, 可以预先定义一些通用的查询概念. 利用上下文知识, 查询概念可映射到基于服务器和基于客户端的具体查询上. 查询的结果文档被重新改写, 以便引入一些指向有用信息的链. 查询概念的转换及结果文档的改写是基于公共网关接口的(Common Gateway Interface, CGI) [5]. 一个 CGI 应用将概念化查询映射到适当的关键字. 在这一步查询概念被转换成查询关键字的逻辑运算. 在结果返回给客户之前, 所有被引用的文档上的链被重新改写, 以便描述另一个局部CGI 应用, 原始的 URL 和查询上下文(概念和实例) 作为该应用的参数.

3.2 知识库

通常, 信息检索本身不是目的. 每个人总是带着一定的目的去查询信息, 并希望以此达到目的. 如果系统知道用户的目标, 就可以帮助他更好地描述他的查询, 知识库存放的就是这些相关信息.

规划识别(Plan Recognition) [4] 和用户建模可用以从用户的动作推导出他的目的.

· 规划识别: 利用一组预定义的规划. 系统通过观察用户的动作来预测用户下一步的动作.

· 用户模型是描述用户关于某一特殊应用的长期行为, 即他的目标、领域知识和他的概貌. 通过这些信息, 系统可预测用户的信息需求.

当用户在不同时间有不同的目标时, 用户建模方法可能产生过多的信息. 另外, 当一个公司的若干雇员从事相同的工作, 有相同的目标, 但用户建模方法可能造成相同目标的冗余. 为解决这些问题, 有必要区分三种不同的知识: 用户模型、任务模型和领域知识.

有许多工具可用于为企业的组织结构和商务处理过程建模. 这些模型稍加扩充就可成为用户模型和任务模型. 在组织结构中, 它描述了每个部门及其作用, 每个雇员及其职责. 其中, 每个雇员的描述就是一个精确的用户模型. 与用户模型不同, 任务模型是描述用户检索信息想要达到的目的. 我们的任务模型可以是面向商务处理过程的. 一个商务处理模型由一组活动, 活动间的关系以及数据流的表示三部分组成 [5]. 对模型中每一个活动, 我们赋上一个或多个查询概念, 用以描述执行这些任务有关的信息.

领域知识包括一个面向应用的词库, 它包括领域中常用词及其同义词的描述; 词与词之间的关系等等. 领域知识描述了一些常识性知识, 它为信息检索提供灵活有效的支持.

4 查询概念及其转换

检索工具定期扫描网络并利用信息检索技术形成基于文本的索引. 这些索引可用以处理用户提出的查询要求. 但检索工具不总能对文档的某些部分进行检索, 其中包括段落或句子等. 这是由于信息检索技术通常是这样操作的: 当一个文档被索引, 就对索引项的出现次数在全文范围内进行计算. 这些结果向量用于对文档建档以便以后检索文档时使用. 为使检索结果满足用户的要求, 有必要对文档的建档技术进行改进, 但这将导致大量的空间开销和过长的访问时间. 另一种方法就是进一步分析文档与用户要求间的关系, 建立有用的连接, 这里, 我们采用第二种方法.

我们提供查询概念来初始化一个检索过程. 查询概念是一个可推导出基本查询的信息组成的集合. 这些查询概念隐含的知识是多方面的:

• 参数: 它刻划概念的动态特性. 一个定义好的关键字和模式具有通用性, 它可包含一些用于实例化每个单独任务的参数.

• 查询串: 描述那些将提交给远程检索工具的字符串. 一个查询可包含变量、函数和逻辑运算. 函数 syn(string) 返回串 string 和它在词库中定义的同义词.

• 限定模式: 描述通用模式集, 它们将作为局部分析的输入. 这些定义不仅要保留常规的模式, 还要保留 HTML 环境, 如 title, header 等, 函数 qual(pattern) 将某一相关文档赋给一个模式, idx(pattern) 说明必须建立索引的位置. 它将在文档改写中被引用. 关键字 'IN' 是一个范围限定符, 它是一个表达式列表

用于说明必须考虑的 HTML 环境. 同样地, SEQ 也是定义范围, 它只有与前面的范围限定符联用才有效.

- *喜欢的位置*: 描述检索工具缺省的查询网页.
- *附加的方法*: 作为一个可选项, 它用以启动附加的分析工具来定位期望的信息.

例如, 假设我们想购买一台计算机, 通过网页检索能够得到计算机的有关性能参数和价格. 这一概念可描述如图 2 所示.

```
Concept: "Product-Comparison"
Parameter: < Product-Name>
Query      : syn(< Product-Name> ) AND
           : syn( "specification ") AND
           : syn( "price ")
Pattern    : qual( idx (
                 syn( "specification ") OR syn ( "price "))
                 IN "< H[1- 4]> " SEQ
                 idx (syn(< Product-Name> )))
```

图 2　一个查询概念的例子

这个例子定义一个 Product-Comparison 概念, 它有一个参数 Product-Name, 这里赋值为 Computer. 查询属性定义了产品名 "Product-Name", 技术指标 "specification" 和价格 "price" 以及它们的同义词的与操作. 模式定义说明在 HTML 头环境 1 至 4 级(即< H [1- 4]>) 找 "specification" 和 "price" 及其同义词, 并且在它们的后继段落中有产品名< Poduct-Name> 的实例或其同义词出现.

当用户开始信息检索, 他说明的查询属性被转换成一系列关键字的逻辑运算, 它们可被已有的检索工具所理解. 因此, 查询中的变量由它们的实例所代替. 系统也可要求用户选择希望的检索工具, 这项说明可在概念定义上加以说明. 被转换后的查询最终被送到适当的检索工具上进行检索.

5　文档分析与改写

在查询概念转换成基本的查询关键字的同时, 所选择的概念和变量实例也将保留下来以便对返回文档做分析和改写.

文档分析主要是借助于模式匹配器进行的. 它是在传统的串匹配器的基础上增加 HTML 文档的处理功能而形成的, 因此具有以下特点:

(1) 模式匹配器在一定范围内围绕某一关键字查找一组匹配的字符串. 在一定范围内关键字/ 字符串的结合称为字模式;

(2) 一个模式也可精确地说明剔除某些字符串;

(3) 它支持近似匹配功能以便处理键入的字母错误;

(4) 它可以识别由 HTML 标识符表明的某些环境;

(5) 基于在文本中匹配的模式, 它可以为文档计算一个统计的相关值;

(6) 它可用命名锚(named anchor) 标注在文档中匹配的模式, 并产生一个指向这些位置的索引.

局部分析首先是用于选择感兴趣的文档, 出现的模式本身就是一个有用的信息. 因此在文档适当的部分标注上命名锚以便允许直接访问它们. 指向这些部分的链可形成一个索引附加到文档的末尾.

图 3 给出一个例子说明一个文档和改写后的结果文档. 其中关键字 price 和变量 Product-Name 的值 computer 在文档中分别出现一次, 因此产生的索引包含两项, 它们分别指向检索关键字出现的位置.

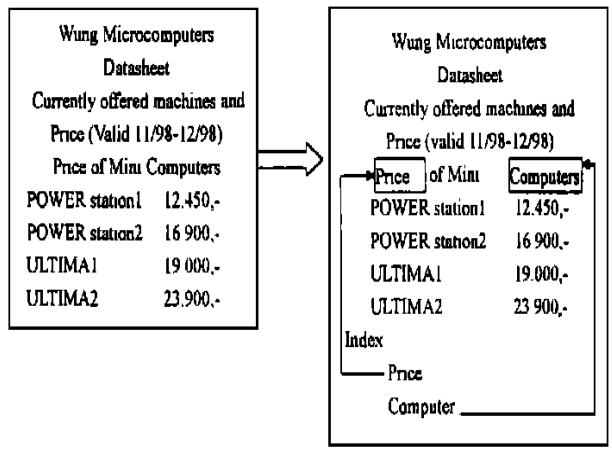

图 3　文档改写

结构信息可用于控制和影响相关的分析过程以及加强信息检索机制的功能. 对于大量的 HTML 文档, 它的结构信息很容易得到. 但对那些纯 ascii 文档(如 Gopher、Net-News 和 Email 当等) , 它们没有直接的标注. 我们可利用一些公共的启发式信息, 如首行空格, 列表符号或枚举符号, 空行等等将一个文档转换成一个 HTML 文档. 转换后的 HTML 文档可被模式匹配器分析.

6　结束语

基于知识的网页检索工具 KWSE 是将人工智能的基本思想与信息检索技术相结合的产物, 它避免了基本关键字检索方法所造成的信息冗余和有用信息的丢失, 因此具有广阔的应用前景. KWSE 虽然仍处在研制阶段, 但一些关键技术的实验表明, KWSE 的检索结果更为有效, 更符合人们的思维习惯.

该系统的下一步研究工作是考虑如何将它与企业知识管理系[6]的结合. 企业知识管理系统是一门新兴的交叉学科, 主要研究如何形式化地管理知识资源, 以便在企业范围内有效地访问和使用知识. 因为在某一特定的应用背景中, KWSE 更能发挥其特长.

参　考　文　献

1　A. Aamodt, M. Nygard. Different roles and mutual dependencies of data, information, and knowledge-An AI Perspective on Their Integration. [J] In Data & Knowledge Engineering 16 (1995) Elsevier, North-Holland, 191～222

2　M. Erdmann. The data warehouse as a means to support

knowledge management. [C] In Workshop of Knowledge-Based System for Knowledge Management in Enterprise in KI97, Sept. 1997, Germany
3 The Common Gateway Interface, http://hoohoo.ncsa.uiuc.edu/cgi/
4 Jana Roehler. Planning from second principles-a logic-based approach. [R] Research Report. RR-94-13, DFKI-German Research Center for Artificial Intelligence. 1994
5 Dimitris Karagiannis, (ed.) ACM SIGOIS Bulletin. Special issue: business process reengineering. [J] ACM Special Interest Group on Office Information Systems. 1995
6 D. O'Leary. Enterprise knowledge management. [J] IEEE Computer, 31(3): 54-61, March 1998, 54～61

A KNOWLEDGE-BASED WEB SEARCH ENGINE

LIAO Ming-hong WU Xiang-hu
(*Dept. of Computer Science of Harbin Institute of Technology Harbin* 150001)

Abstract With the world widely using of the Internet, more and more people do their research work and business activities based on the Internet, and search engines become indispensable tools to them. However, the currently common search engines are based on keywords query, which usually causes the overloading of information or losing of useful information. There are two reasons about it: one is that the query provided by the user cannot fully describe his goal; and the other is that there is not effective index in the result documents, which guides users to their useful information. So we provides a knowledge-based web search engine(KWSE), which expands query concept and document rewriting in the common search engine, i.e. users submit a concept with some goal to the search engine, instead of concrete keywords, and the result document is appended with some indexes which point to the useful information. Some expriments indicate, KWSE can produce more effective results.

Key words Internet; WWW; Information retrieval